MID-RATE CODING BASED ON A SINUSOIDAL REPRESENTATION OF SPEECH

Robert J. McAulay and Thomas F. Quatieri

Lincoln Laboratory, Massachusetts Institute of Technology
Lexington, Massachusetts 02173-0073

ABSTRACT

In this paper a sinusoidal model for the speech waveform is used to develop a new analysis/synthesis technique that is characterized by the amplitudes, frequencies, and phases of the component sine waves. The resulting synthetic waveform preserves the waveform shape and is essentially perceptually indistinguishable from the original speech. Furthermore, in the presence of noise the perceptual characteristics of the speech and the noise are maintained. Based on this system, a coder operating at 8 kbps is developed that codes the amplitudes and phases of each of the sine wave components and uses a harmonic model to code all of the frequencies. Since not all of the phases can be coded, a high frequency regeneration technique is developed that exploits the properties of the sinusoidal representation of the coded baseband signal. Based on a relatively limited data base, computer simulation has demonstrated that coded speech of good quality can be achieved. A real-time simulation is being developed to provide a more thorough evaluation of the algorithm.

INTRODUCTION

In a previous paper by the authors [1], a "magnitude-only" analysis/synthesis system was developed based on the intuition that the speech waveform could be characterized by the amplitudes and frequencies of an underlying set of sine waves. These parameters were estimated by locating the peaks of the magnitude of the high-resolution short-time Fourier transform. In order to set up amplitude and phase tracks that could be applied to a bank of sine wave generators, it was necessary to match the peaks obtained on contiguous frames. This was done using a nearest neighbor frequency tracker that allowed for the "birth" and "death" of sine waves. This capability was essential for allowing the tracker to adapt to rapidly varying speech events due to pitch variations, and to changes in the voicing state. With the establishment of continuous frequency tracks the sine wave phase was defined to be the integral of the instantaneous frequency computed along each track.

This system yielded synthetic speech that was quite good. Based on this system, a speech coder was developed that operated at 8 kbps and produced acceptable synthetic speech that was noticably free of artifacts. The major weakness of both the uncoded and coded systems occurred for low-pitch speakers and for speech in noise. In the former case, the synthetic speech was of good quality but it was perceptually different from the original. In the latter case the synthetic noise took on a tonal quality that was unnatural and annoying. This quality rendered the system unacceptable for applications where robust synthetic speech was required.

Since the magnitude-only system made use of only the amplitudes and frequencies of the underlying sine waves, then if further improvements were to be obtained using the sinusoidal speech model, they could only result from the inclusion of the measurements of the sine wave phases. In this paper an algorithm for unwrapping and interpolating the sine wave phases is developed that produces a phase track that is consistent with the measured frequencies and phases. When these phase functions are used in the sine wave generators, synthetic speech is produced that obviates the problems encountered with the magnitude-only system. Then by coding the phases it becomes possible to produce synthetic speech at 8 kbps that is of good quality even when the speech is in a background of noise.

THE SINUSOIDAL SPEECH MODEL

Following the previous work by the authors [1], the speech waveform is modelled as the sum of sine waves. If s(n) represents the speech waveform, then

$$s(n) = \sum_{\ell=1}^{L} A_\ell(n) \sin[\theta_\ell(n)] \qquad (1)$$

where $A_\ell(n)$ and $\theta_\ell(n)$ are the time-varying amplitudes and phases of the ℓ'th tone. In order to determine the amplitude and phase functions, a high-resolution Discrete Fourier Transform (DFT) is computed every frame (10-20 ms) and a set of sine wave frequencies is generated by applying simple peak-picking to the magnitude function. The estimates of the parameters to be used in generating (1) correspond to the amplitudes, frequencies and phases that are measured from the DFT at the locations of the peaks.

In order to set up amplitude and phase tracks for the sine wave generators, it is necessary to match the parameters obtained on contiguous frames. This can be done using the nearest neighbor frequency tracking algorithm that was developed for the magnitude-only analysis/synthesis system [1]. As a result of this frequency matching algorithm, all of the parameters measured for an arbitrary frame k are associated with a corresponding set of parameters for frame k+1. Letting $(\hat{A}^k_\ell, \hat{\omega}^k_\ell, \hat{\theta}^k_\ell)$ and $(\hat{A}^{k+1}_\ell, \hat{\omega}^{k+1}_\ell, \hat{\theta}^{k+1}_\ell)$ denote the successive sets of parameters for the ℓ'th

CH2118-8/85/0000-0945 $1.00 © 1985 IEEE

frequency track, then an obvious solution to the amplitude interpolation problem is to take

$$\tilde{A}(n) = \hat{A}^k + \frac{(\hat{A}^{k+1} - \hat{A}^k)}{N} n \qquad (2)$$

where $\tilde{A}(n)$ denotes the estimate of the amplitude to be used in (1) and n = 1, 2,..., N is the time sample into the k'th frame. (The track subscript "ℓ" has been omitted for convenience).

Unfortunately such a simple approach cannot be used to interpolate the frequency and phase because the measured phase, $\hat{\theta}^k$, is obtained modulo 2π. Hence, phase unwrapping must be performed to insure that the frequency tracks are "maximally smooth" across frame boundaries, a concept which will be made clear in the sequel. The first step in solving this problem is to postulate a phase interpolation function that is a cubic polynomial, namely

$$\theta(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3 \qquad (3)$$

It is convenient to treat the phase function as though it were a function of a continuous time variable t, with t=0 corresponding to frame k and t=T corresponding to frame k+1. The parameters of the polynomial must be chosen to satisfy the frequency and phase measurements obtained at the frame boundaries. Since the instantaneous frequency is the derivative of the phase, then

$$\dot{\theta}(t) = \gamma + 2\ \alpha t + 3\ \beta t^2 \qquad (4)$$

and it follows that at the starting point t=0,

$$\begin{aligned} \theta(0) &= \zeta = \hat{\theta}^k \\ \dot{\theta}(0) &= \gamma = \hat{\omega}^k \end{aligned} \qquad (5)$$

and at the terminal point t=T

$$\begin{aligned} \theta(T) &= \hat{\theta}^k + \hat{\omega}^k T + \alpha T^2 + \beta T^3 = \hat{\theta}^{k+1} + 2\pi M \\ \dot{\theta}(T) &= \hat{\omega}^k + 2\alpha T + 3\beta T^2 = \hat{\omega}^{k+1} \end{aligned} \qquad (6)$$

where again the track subscript "ℓ" is omitted for convenience.

Since the terminal phase $\hat{\theta}^{k+1}$ is measured modulo 2π, it is necessary to augment it by the term $2\pi M$ (M is an integer) in order to make the resulting frequency function "maximally smooth". At this point M is unknown, but for each value of M, whatever it may be, (6) can be solved for $\alpha(M)$ and $\beta(M)$, (the dependence on M has now been made explicit). The solution is easily shown to satisfy the matrix equation

$$\begin{bmatrix} \alpha(M) \\ \beta(M) \end{bmatrix} = \begin{bmatrix} \frac{3}{T^2} & \frac{-1}{T} \\ \frac{-2}{T^3} & \frac{1}{T^2} \end{bmatrix} \begin{bmatrix} \hat{\theta}^{k+1} - \hat{\theta}^k - \hat{\omega}^k T + 2\pi M \\ \hat{\omega}^{k+1} - \hat{\omega}^k \end{bmatrix} . \qquad (7)$$

In order to determine M and ultimately the solution to the phase unwrapping problem, an additional constraint needs to be imposed that quantifies the "maximally smooth" criterion. Figure 1 illustrates a typical set of cubic phase interpolation functions for a number of values of M. It seems clear on intuitive grounds that the best phase function to pick is the one that would have the least variation. In Fig. 1, for example, the "best" phase trajectory would be the cubic function corresponding to M=2. This is what is meant by a maximally smooth phase track. In fact, if the frequencies were constant and the vocal tract were stationary, the true phase would be linear. Therefore a reasonable criterion for "smoothness" is to choose M such that

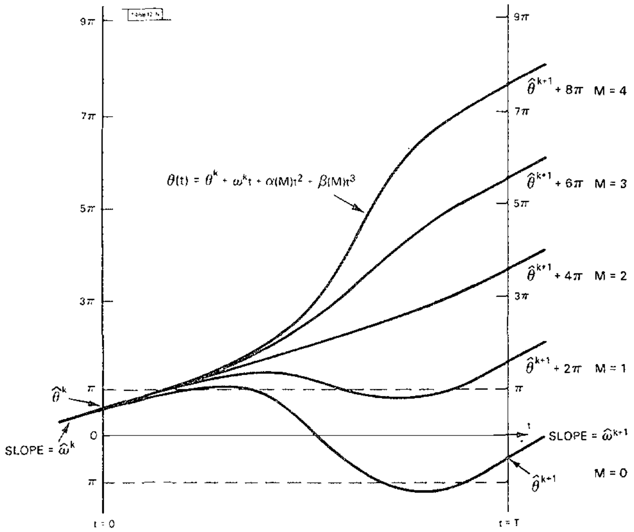

Fig. 1. A typical set of cubic phase interpolation functions.

$$f(M) = \int_0^T [\ddot{\theta}(t;M)]^2\, dt \qquad (8)$$

is a minimum, where $\ddot{\theta}(t;M)$ denotes second derivative of $\theta(t;M)$ with respect to the time variable, t.

Although M is integer valued, since f(M) is quadratic in M, the problem is most easily solved by minimizing f(x) with respect to the continuous variable x and then choosing M to be the integer closest to x. After straightforward but tedious algebra, it can be shown that the minimizing value of x is

$$x^* = \frac{1}{2\pi} \left[(\tilde{\theta}^k + \tilde{\omega}^k T - \hat{\theta}^{k+1}) + (\hat{\omega}^{k+1} - \hat{\omega}^k) \frac{T}{2} \right] \qquad (9)$$

from which M* is determined and used in (7) to compute $\alpha(M^*)$ and $\beta(M^*)$, and in turn, the unwrapped phase interpolation function

$$\theta(t) = \hat{\theta}^k + \hat{\omega}^k t + \alpha(M^*)t^2 + \beta(M^*)t^3 \qquad (10)$$

This phase function not only satisfies all of the measured phase and frequency endpoint constraints, but also unwraps the phase in such a way that $\theta(t)$ is maximally smooth.

Since the above analysis began with the assumption of an initial unwrapped phase $\hat{\theta}^k$ corresponding to frequency $\hat{\omega}^k$ at the start of frame k, it is necessary to specify the initialization of the frame interpolation procedure. This is done by noting that at some point in time the track under study was born. When this event occurred, an amplitude, frequency and phase were measured at frame k+1 and the parameters at frame k to which these measurements correspond were defined by setting the amplitude to zero (i.e., $\hat{A}^k = 0$) while maintaining the same frequency (i.e., $\hat{\omega}^k = \hat{\omega}^{k+1}$). In order to insure that the phase interpolation constraints are satisfied initially, the unwrapped phase $\hat{\theta}^{k+1}$ is defined to be the measured phase and the start-up phase is defined to be

$$\hat{\theta}^k = \hat{\theta}^{k+1} - \hat{\omega}^{k+1} N \qquad (11)$$

where N is the number of samples traversed in going from frame k+1 back to frame k.

Letting $\tilde{\theta}_\ell(t)$ denote the unwrapped phase function for the ℓ'th track obtained from the above procedure, then the final synthetic waveform is given by

$$\tilde{s}(n) = \sum_{\ell=1}^{L^k} \tilde{A}_\ell(n) \cos\,[\tilde{\theta}_\ell(n)] \qquad (12)$$

where $kN < n < (k+1)N$, $\tilde{A}_\ell(n)$ is given by (2), $\tilde{\theta}_\ell(n)$ is the sampled data version of (10), and L^k is the number of sine waves estimated for the k'th frame.

As a result of the above phase unwrapping procedure, each frequency track will have associated with it an instantaneous unwrapped phase which accounts for both the rapid phase changes due to the frequency of each sinusoidal component, and the slowly varying phase changes due to the glottal pulse and the vocal track transfer function. A more refined model has been developed that separates the effects of the excitation phase from the vocal tract phase [2]. Explicitly, separating the phases in this way is useful in time-scale modification as it allows for independent control of the rate of articulation of the vocal tract.

EXPERIMENTAL RESULTS

A block diagram description of the complete analysis/synthesis system is given in Fig. 2. A non-real time floating point simulation was developed in order to determine the effectiveness of the proposed approach in modelling real speech. The speech processed in the simulation was low-pass filtered at 5 kHz, digitized at 10 kHz and analyzed at 10 ms frame intervals. A 512-point DFT using a 20-30 ms Hamming window was found to be sufficient for accurate peak estimation. The maximum number of peaks that could be specified was limited by a threshold that was a function of the average pitch. In general the performance was little affected by the choice of this threshold unless, of course, too few peaks were allowed. Although only a relatively small amount of speech data has been processed (≈ 5 minutes), in all the cases studied the synthetic speech was essentially perceptually indistinguishable from the original. Visual examination of many of the reconstructed passages shows that the waveform structure is essentially preserved. Hence the quasi-stationarity conditions seem to be satisfactorily met and the use of the parametric model based on the amplitudes, frequencies and phases of a set of sine wave components appears to be justifiable.

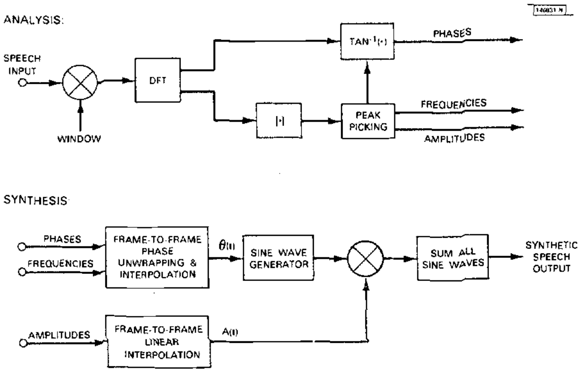

Fig. 2. Block diagram of sinusoidal analysis/synthesis system.

When tested for low-pitched speakers and for speech in noise the same desirable properties were found to hold and, in particular, the synthetic noise had the same perceptual qualities as the original noise. Therefore the inclusion of measured phase via the cubic unwrapping and interpolation algorithm has overcome the difficulties inherent in the magnitude-only analysis/synthesis system.

CODING AT 8 KBPS

The original motivation for this investigation was to develop a speech coding system for operation at 8 kbps. It has been shown that the key to synthesizing high-quality speech using the sinusoidal speech model depends critically on the inclusion of the measured phases of each sine wave component. Therefore coding the phases becomes of highest priority. Since the sinusoidal representation also requires the specification of the amplitudes and frequencies, it is clear that relatively few peaks could be coded before all of the available bits were used. The first step, therefore, is to significantly reduce the number of parameters that must be coded. One way to do this is to force all of the frequencies to be harmonic.

That this choice should be the one to focus on follows from the fact that during voiced speech one would expect all of the peaks to be harmonically related and therefore, by coding the fundamental, the locations of all of the frequencies will be available at the receiver. It would seem that problems would occur during unvoiced speech since the frequency locations of the peaks will not be harmonic in this case. However, it is well known from random process theory [3] that noise-like waveforms can be represented (in an ensemble mean-squared error sense) in terms of a harmonic expansion of sine waves provided the spacing between adjacent harmonics is small enough that there is little change in the power spectrum envelope (i.e. intervals less than about 100 Hz). This representation preserves the statistical properties of the input speech provided the amplitudes and phases are randomly varying from frame to frame. Since the amplitudes and phasesare to be coded, this random variation inherent in the measurement variables can be preserved in the synthetic waveform.

As a practical matter it will be necessary to estimate the fundamental frequency that characterizes the set of frequencies in each frame, which in turn relates to pitch extraction. Suffice it to say that a new frequency-based pitch extraction algorithm has been developed that selects the fundamental frequency of a harmonic set of sine waves to produce the best fit to the input waveform according to a perceptual criterion. This new algorithm is described in detail in a report by the authors [4].

As an immediate consequence of using the harmonic frequency model, it follows that the number of sine wave components to be coded is the bandwidth of the coded speech divided by the fundamental. Since there is no guarantee that the number of measured peaks will equal this harmonic number, provision must be made for adjusting the number of peaks to be coded. Based on the fundamental, a set of harmonic frequency bins are established and the number of peaks falling within each bin are examined. If more than one peak is found, then only the amplitude and phase corresponding to the largest peak are retained for coding. If there are no peaks in a given bin, then a fictitious peak is created having an amplitude and phase obtained by sampling the short-time Fourier Transform at the frequency corresponding to the center of the bin.

The amplitudes are then coded by applying the same techniques used in the channel vocoder [5]. That is, a gain level is set by using 5 bits with 2 dB per level to code the peak of the fundamental. Subsequent peaks are coded logarithmically using delta-modulation techniques across frequency. In the simulation 3.6 kbps are assigned to code the amplitudes at a 50 Hz frame rate. Adaptive bit allocation rules are used to assign bits to peaks. For example, if the pitch is high there will be relatively few peaks to code, and there will be more bits per peak. Conversely when the pitch is low there will be relatively few bits per peak, but since the peaks will be closer together their values will be more correlated, hence the ADPCM coder should be able to track them well.

To code the phases a fixed number of bits per peak (typically 4 or 5) is used. Since there remains only 4.4 kbps to code the phases and the fundamental (7 bits are used), then at a 50 Hz frame rate, it will be possible to code at most 16 peaks. At a 4 kHz speech bandwidth, all of the phases will be coded provided the pitch is greater than 250 Hz. If the pitch is less than 250 Hz provision has to be made for regenerating a phase track for the uncoded high frequency peaks. This is done by translating the instantaneous frequency track obtained from the cubic interpolation function for the low frequency coded peaks to the frequencies of the uncoded peaks. In this way the phase coherence intrinsic to voiced speech and the phase incoherence characteristic of unvoiced speech is effectively translated to the uncoded frequency regions.

Using a non-real-time floating point computer simulation very high quality synthetic speech has been obtained. Moreover the quality was maintained for noisy speech, preserving not only the quality of the original speech but also of the noise as well. Therefore the method appears to be very attractive for the robust mid-rate coding applications. To more thoroughly evaluate the algorithm, a real-time system is being developed on the Lincoln Digital Signal Processor [6].

CONCLUSIONS

Motivated by the need to obtain a high-quality robust speech coder at 8 kbps some preliminary experiments based on a sinusoidal model for the speech waveform were performed. It was found that to maintain high-quality for low-pitched speakers and for speech in noise it was necessary to include the measured phase for each sine wave component. This was done using a cubic polynomial to unwrap the phase and provide a continuous interpolation function between contiguous frames.

The parameters of the resulting analysis/synthesis system were coded at 8 kbps by using a harmonic sine wave model based upon an estimated fundamental that provided the best "perceptual" fit to the input data. Channel vocoder techniques were used to code the amplitudes at 3.6 kbps. The remaining 4.4 kbps were used to code the fundamental and the measured phases. For lower pitched speakers, not all of the phases could be coded, hence a high frequency regeneration technique was developed that exploited the properties of the sinusoidal representation of the baseband signal. In general, high-quality coded speech was obtained even for speech in noise, but final determination of the efficacy of the algorithm will be deferred until a real-time simulation has been completed.

REFERENCES

[1] R.J. McAulay and T.F. Quatieri, "Magnitude-Only Reconstruction Using a Sinusoidal Speech Model," ICASSP '84, International Conference on Acoustics, Speech and Signal Processing, San Diego, CA, March 19-21, 1984, pp. 27.6.1 - 27.6.4.

[2] T.F. Quatieri and R.J. McAulay, "Speech Transformations Based on a Sinusoidal Representation," ICASSP '85, International Conference on Acoustics, Speech and Signal Processing, Tampa, FL, March 26-29, this proceedings.

[3] H. Van Trees, Detection Estimation and Modulation Theory, Part I, Wiley, New York, 1968, Chapter 3.

[4] R.J. McAulay and T.F. Quatieri, "A Sinusoidal Representation of Speech, Volume III: Speech Coding at 8 kbps," M.I.T. Lincoln Laboratory Technical Report, TR-693, January 1985.

[5] J.N. Holmes, "The JSRU Channel Vocoder," IEEE Proc. Vol. 127, Pt. F., No. 1, February 1980, pp. 53-60.

[6] P.E. Blankenship, "LDVT: High Performance Minicomputer for Real-Time Speech Processing," in EASCON '77, Rec., September 1977.

*This work was sponsored by the Department of the Air Force.

The U.S. Government assumes no responsibility for the information presented.